What is Data Science?

Welcome to this course on introduction to data science.
So before we get started I wanted to just kinda give you a bit about my background.
So, my classical training is in molecular biology.
I did my graduate work at UC San Diego.
I worked at the Salk Institute, and it was a really exciting time in molecular biology, so this was in the early 2000's.
Big genome sequencing was just starting to kind of take off.
Craig Venter, when I started,
had just finished sequencing the human genome.
People at the Salk Institute, Joe Ecker in particular, was sequencing mRNAs and looking at global gene expression sorts of data.
Where we can look at all of the potential proteins that were going to be produced by an organism at a certain time point.
And so there's this huge amount of data coming in into biology.
And at the time I was kind of peripherally involved in that.
I was working on specific genes,
specific protein-protein interactions.
But I was right next to all these exciting stuff going on, and it really started making me think about data science and how do you deal with these big data structures and in the process I kind of was stuck in that kind of white castle of academia and focused on my own work.
But it's been influencing me and sitting in the back of my mind for years now.
And after I finished my graduate work and did a post doc working on algae biofuels which turned out to be an economically not viable attempt to kind of solve our fuel problems at current juncture.
I decided to move into computer science and kind of remove that black box of biology and move into these silver boxes of computers.
And that transition really has allowed me to kinda dig in
to the data science and computer science and
really think about how to attack those big pieces of data.
And so the goal of this course isn't really to show you all
the amazing things you can do with data science but
really talk about kind of a big picture of data science.
And what is data science?
I think a lot of people are kind of unclear as to what data science is, and then from there kinda give you some basic tools to kind of attack some data science yourself.
And we'll talk about why you wanna have those basic tools in place, and kind of we can go from there.
So with that, let's go ahead and start of and
talk a little bit about data science and what is data science.
And then throughout the rest of the program we're gonna talk about data

organization in our next chapter, data visualization, data analytics,
how to become a data scientist and then finally as I said we're gonna
give a little bit of demo on using Excel.
All right so what is data science?
So there're kinda two major pieces of data science,
the first on is this idea of just data analysis.
And so a lot of jobs and companies that are looking for data scientist,
they are really looking for somebody that's an expert in data analysis.
So this somebody that can take data and
with that data they can make these really cool inferences or
extrapolate information that's valuable to the company.
So that's kind of one side of data science.
But there's kind of a second side of data science and
that's really the scientific approach to data itself.
And this would be what sort of mechanisms or
strategies are used to best attack a dataset.
So data is just data, right, so it doesn't matter if it comes from
a biological system, or physical system, or from internet use or
wherever the data comes from it doesn't make a difference.
It's really about the fact that is data and have certain behaviors that
one can apply scientific methods effectively to it.
And so that's kind of the second part of data science.
So data analysis is kind of one aspect of where it's like,
hey I'm gonna use these tools to really create a really good
predictive model or whatever.
Or, and I should say, there's the second type,
which is kind of this more academic structure.
Which is like, how do I best attack data?
So do I use clustering?
Do I use a neural net?
Do I use a support vector machine to attack this data in the most
effective way.
And so there's that kind of split in what a data scientist is.
And so, I think that generates some of the nebulousness of, hey,
what is a data scientist?
And so, it's kinda these both of pieces that kinda have been munged
into and collapsed into one thing.
And so how did data science really kind of get its speed under itself?
So it started of kind of as a potentially as an evolution of
applied statistics, but
you could also think of it as an evolution of computer Science.
So if we look at the Scientific Method, kinda the general plan
with Scientific Method is you have data, right?
And you're looking at that data and you wanna understand it, right?
Scientists, the goal of the scientists is to understand
how things work right.

So you are gonna question about that data.
So you have this question in mind and
you become hypothesized why that happens.
So, you have this kind of model which is this hypothesis of like,
hey, the answer to my question is blah.
So, then you set up an experiment.
Then you analyze the result of that experiment.
And then you either say, hey, this is a result, like the data so
far fit this hypothesis.
So I'm gonna kinda present that as like hey, my hypothesis is working.
I might have generated new data that allows me to subtly mutate or
change my hypothesis.
Or it would be like hey, this is a really cool result, let's talk and
share it with the world.
But the majority of the time what happens is you get your data back
and you're like that doesn't really fit my hypothesis, or
doesn't fit it as well as I would like.
And so you end up with this kind of iteration and this iterative cycle.
So you basically, you hypothesis, experiment,
analyze, hypothesize again, experiment, analyze and repeat.
And so one of the slowest points of that
process in some cases is the analysis of that data right.
So you get all this data back and now you have to do all these
analyses to see hey does it actually fit my hypothesis.
And so you might have at the time 10,000 points of data or
even a 100,000 points of data.
And if you had to do that by hand, or
even write your own computer program to do it that writing and
doing by my hand just takes a lot of time.
And so, computers came in said hey, we can well,
computer didn't say anything but we can computers to do that analysis.
And so,
that's kinda the first chunk of data science is doing that analysis.
But over time it started to evolve where not only do we
need to do the analysis or can we do the analysis,
we can also have it do different hypotheses.
And so an example of this might be, hey,
we're gonna predict that two events so let's just call them x and
y fit together and correlate with a straight line.
And so we generate that hypothesis and we try to draw a whole bunch of
lines, mx + b sort of lines where these straight lines and does that
fit these data points and what the best fit of that data point?
But that's great.
But you could maybe fit those data better with a polynomial.
So some sort of curved line.
And every time you change that line shape,

you have to change all this analysis.
And you're basically running a different hypothesis on that data.
You're saying, hey this line fits best, or this Polynomial line,
fits best.
Or this logarithmic line fits best.
And so each one of those take a lot of time.
And so you as a person would have to come up with those different hypotheses and run that experiment and then do the data analysis, and hopefully the computer would do the data analysis for you.
But people realized that there's no reason that
you have to come up with that hypothesis each time.
You could run a suite of hypothesis and basically computer science has start to evolve where it's basically doing the hypothesis,
the experiment and then the analysis kind of that whole chunk.
And some of the really cool AI stuff is actually even started at kinda move out and take, And
really attack the questions that are there.
So it's not just doing the hypothesis experiment
analysis side of the kind of the scientific method.
We're moving into having those machines actually ask the question and look at the data, and kinda taking this whole loop.
And so that's kind of the future I think of some pieces of data
science and AI.
Awesome, so what have some of the successes of data science been?
So obviously Google.
So search engine optimization has been a huge hit for
a number of companies.
And the reason it's ahead is basically these companies are taking
in this tremendous amount of data about someone right.
We're generating all this what's called data exhaust as we go
through the world what we search on the Internet all those information pieces what ads we look at etc all get collected by someone.
And because those data are collected you can analyze that data and generate, hey maybe this person's gonna buy this product
in the next year.
So maybe we wanna send them a few extra ads in that area.
And basically Google said, hey let's take that data,
let's sell that data basically to bring in more ad revenue.
And that was kind of their basic general model.
But basically that was all of this kind of integration of
let's make some we had this data let's make a hypothesis about it
let's experiment let's analyze it.
And see what kind of best ads we can send to people that
have these particular sets of behaviors.
But other things they've been successful.
Amazon has basically gone through all their product catalog and

have an algorithm that goes through make sure that the prices don't have any weird price anomalies.
So that they're actually able to make sure that they have a certain margin.
Even if there's a small loss in some cases, they can minimise that margin.
And so they don't have these, every once in awhile they'll have these weird price anomalies that they used to, UPS, when it's kinda disc, right turn algorithms.
So if you haven't heard about this UPS and I think most shipping companies spend the majority of their money on fuel.
And so if you can minimize how much fuel you're using as a driver, you can maximize your cost.
And so they basically applied learning algorithm to driving routes to kind of minimize the amount of fuel that their drivers is gonna be using and so, again, they kind of using this iterative attack on data.
So data science as we kinda mentioned has solved a bunch of kind of interesting problems but it's used all over the place.
So small businesses can benefit from good data analysis and good data scientists you could have great information coming out in genomics.
We already talked about search engine optimization.
Literature, looking at historical literature and looking at what kind of verbage was being used can be really interesting.
So there's all these kinds of really kind of some of them are asterisk, some are really, Apple pod to the times, politics, all these things are being attached from this kind of data-rich environment.
And so we need all these to tools to kind of really expand our knowledge about these different things.
And so, that's why data science is so exciting right now.
So, with that said, hopefully you kinda have general sense of what it does, scientist is so,but just to reiterate.
So a data scientist is someone who either can think about data and kind of the approach to data and how to analyze data in general.
It's kind of academic side of it.
But it's also taking the tools that that academic side are generating and applying them to different types of data.
And be like so I'm gonna use the support vector machine to figure out how these things cluster together.
And then that will lead to different outcomes.
So how do you become a data scientist?
I think there are three major pieces.
So the first piece is organizing data.
Data come in whole bunch of forms, whether spreadsheet or streaming video.
You really need to be able to take that data and put in to some format that your system can deal with.

If your system is just you, that's fine.
I mean, you as a person can be a data scientist.
The next piece is visualization of data.
So being able to kind of look at data and see it yourself.
Whether that's a visual representation of data or
some maybe some other type of representation of data.
And also being able to present that data in a visual format that
somebody is not as familiar with data can understand.
I think those are two critical basis so,
visualization is kinda split in to two pieces.
And the last piece is data analysis.
And this is the kinda fun part of data, being a data scientist is like
how do you attack your data such that you understand it better and
you can kinda really dig into it and get that really insightful
gold nugget that's often hiding in the data?
Essentially be like that's a really interesting hypothesis
that's a really interesting result that comes out of this data maybe
we should move our company in this direction.
So those are kinda the three pieces of data scientists of being a data
scientist and that's what we're gonna be talking about in the next
couple of chapters so thank you.